Chapter 2 - A Simple and General Model for Power Analysis

This chapter develops a simple approach to statistical power analysis that is based on the widely-used *F* statistic. The *F* statistic (or some transformation of *F*) is used to test statistical hypotheses in the general linear model (Horton, 1978; Tatsuoka, 1993b), a model that includes all of the variations of correlation and regression analysis (including multiple regression), analysis of variance and covariance (ANOVA and ANCOVA), t-tests for differences in group means, tests of the hypothesis that the effect of treatments takes on a specific value, or a value different from zero. Most of the statistical tests that are used in the social and behavioral sciences can be treated as special cases of this general linear model.

Analyses based on the *F* statistic are not the only approach to statistical power analysis. For example, in the most comprehensive work on power analysis Cohen (1988) constructed power tables for a wide range of statistics and statistical applications, using separate ES measures and power calculations for each class of statistics. Kramer and Thiemann (1987) derived a general model for statistical power analysis based on the intraclass correlation coefficient, and developed methods for evaluating the power of a wide range of test statistics using a single general table based on the intraclass *r*. Lipsey (1990) used the *t*-test as a basis for estimating the statistical power of several statistical tests. Nevertheless, an analytic model based on the *F* statistic has two substantial advantages. First, it can be applied to a wide array of statistical analyses, even those (e.g., correlation coefficients) that do not appear to rely on an *F* distribution. Second, this model can be easily adapted to test a much wider range of hypothesis than the traditional nil hypothesis that the difference between two groups or two treatments, or that the relationship between two variables is exactly zero.

The idea of using the *F* distribution as the basis for a general system of statistical power analysis is hardly an original one; Pearson and Hartley (1951) proposed a similar model over 50 years ago. It is useful, however, to explain the rationale for choosing the *F* distribution in some detail, because the family of statistics based on *F* have a number of characteristics that help to take the mystery out of power analysis.

Basing a model for statistical power analysis on the *F* statistic provides a nice balance between applicability and familiarity. First, the *F* statistic is familiar to most researchers. This chapter and the one that follows show how to transform a wide range of test statistics and measures into *F* statistics, and how to use those *F* values in statistical power analysis. Because such a wide range of statistics can be transformed into *F* values, structuring power analysis around the *F* distribution allows one to cover a great deal of ground with a single set of tables.

Second, the approach to power analysis developed in this chapter is flexible. Unlike other presentations of power analysis, we do not limit ourselves to tests of the traditional null hypothesis (i.e., the hypothesis that treatments have no effect whatsoever). Traditional null hypothesis tests have been roundly criticized (Cohen, 1994; Meehl, 1978; Morrison & Henkel, 1970), and there is a need to move beyond such limited tests. Our discussions of power analysis consider several methods of statistical hypothesis testing, and show how power analysis can be easily extended beyond the traditional null hypothesis test. In particular, we show how the model developed here can be used to evaluate the power of "minimum-effect" hypothesis tests - i.e., tests of the hypothesis that the effects of treatments exceed some pre-determined minimum level.

Recently, researchers have devoted considerable attention to alternatives to traditional null hypothesis tests (e.g., Murphy & Myors, 1999; Rouanet, 1996; Serlin and Lapsley, 1985, 1993), focusing in particular on tests of the hypothesis that the effect of treatments falls within or outside of some range of values. For example, Murphy and

Myors (1999) discuss alternatives to tests of the traditional null hypothesis that involve specifying some range of effects that would be regarded as negligibly small, and then testing the hypothesis that the effect of treatments either falls within this range (H_0) or falls above this range (H_1).

The *F* statistic is particularly well-suited to tests of the hypothesis that effects fall within some range that can be reasonably described as “negligible” vs. falling above that range. The *F* statistic ranges in value from zero to infinity, with larger values indicating stronger effects. As we will show in sections that follow, this property of the *F* statistic makes it easy to adapt familiar testing procedures to evaluate the hypothesis that effects exceed some minimum level, rather than simply evaluating the possibility that treatments have no effect.

Finally, the *F* distribution explicitly incorporates one of the key ideas of statistical power analysis - i.e., that the range of values that might reasonably be expected for a variety of test statistics depends in part on the size of the effect in the population. As we note below, the concept of ES is reflected very nicely in one of the three parameters that determines the distribution of the *F* statistic (i.e., the noncentrality parameter).

The General Linear Model, the F Statistic, and Effect Size

Before exploring the *F* distribution and its use in power analysis, it is useful to describe the key ideas in applying the general linear model as a method of structuring statistical analyses, show how the *F* statistic is used in testing hypotheses according to this model, and describe a very general index of whether treatments have large or small effects.

Suppose 200 children are randomly assigned to one of two methods of reading instruction. Each child receives instruction that is either accompanied by audio-visual aids (computer software that "reads" to the child while showing pictures on a screen) or given without the aids. At the end of the semester each child's reading performance is measured.

One way to structure research on the possible effects of reading instruction is to construct a mathematical model to explain why some children read well and others read poorly. This model might take a simple additive form:

$$y_{ijk} = a_i + b_j + ab_{ij} + e_{ijk} \qquad [1]$$

where: y_{ijk} = the score for child **k**, who received instruction method **i** and audio-visual aid **j**

a_i = the effect of the method of reading instruction

b_j = the effect of audio-visual aids

ab_{ij} = the effect of the interaction between instruction and audio-visual aids

e_{ijk} = the part of the child's score that cannot be explained by the treatments received

When a linear model is used to analyze a study of this sort, researchers can ask several sorts of questions. First, it makes sense to ask whether the effect of a particular treatment or combination of treatments is large enough to rule out sampling error as an explanation for why people receiving one treatment obtain higher scores than people not receiving it. As we explain below, the *F* statistic is well suited for this purpose.

Beginning of Boxed Section

Effect Size

As Preacher and Kelly (2011) note, the term “effect size” is widely used in the social and behavioral sciences, but its meaning is not always clear or consistent. An “effect size” might refer to a sample result (i.e., the sample d is the often referred to as an effect size), to a population parameter (d in the population is unlikely to be the same size as the sample d), or it might refer to a verbal description of a range of effects (e.g., a small effect). They offer a very useful general definition of the term, noting that "Effect size is defined as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelly & Preacher, 2012, p. 140). That is, effect size is not a statistic or a specific parameter, it is any quantitative measure that reflects the magnitude of a phenomenon

Preacher and Kelly (2011) note that effect sizes can be discussed at three different levels of analysis. First, there are the dimensions and units that define particular discussions of effect size. For example in this book, we often use the percentage of variance in one variable explained by some other variable or set of variables as a working definition of effect size. Second, effect size can be discussed in terms of some specific index (e.g., d). Third, they might be discussed in terms of some particular value of that index (e.g., d = .20 is a widely accepted definition of a small difference between two means)

Preacher and Kelly (2011) lay out several criteria for good effect size measures:

- Effect size values should be scaled in a way that makes sense to the reader.
- Effect size values should be accompanied with confidence intervals.
- Your estimate of the population effect size value should be independent of sample size.
- Estimates of effect sizes values should be unbiased (their expected values should equal the corresponding population values), consistent (they should converge to the corresponding population value as sample size increases), and efficient (they should have minimal variance among competing measures).

In this book, we will use the Percentage of Variance (PV) in some variable or outcome that is explained as a general measure of effect size. This is not the only possible effect size measure, and it may not even be the best in some specific circumstances, but it meets the criteria laid out by Preacher and Kelly (2011), and it provides a common language for researchers in different areas to use when discussing their results, even when the statistical methods used (e.g., ANOVA vs. Multiple Regression) appear to differ.

End of Boxed Section

Second, it makes sense to ask whether the effects of treatments are relatively large or relatively small. There are a variety of statistics that might be used in answering this question, but one very general approach is to estimate the percentage of variance in scores (*PV*) that is explained by various effects included in the model (Valentine & Cooper, 2003). We use this measure extensively in this book, in part because it can be easily linked with the widely used F statistic. Regardless of the specific approach taken in statistical testing under the general linear model (e.g., analysis of variance, multiple regression, t-tests), the goal of the model is always to explain variance in the dependent variable (e.g., to help understand why some children obtained higher scores than others).

Linear models like the one shown above divide the total variance in scores into variance that can be explained by methods and treatment effects (i.e., the combined of effects of instruction audio-visual aids) and variance that cannot be explained in terms of the treatments received by subjects. The percentage of variance (*PV*) associated with each effect in a linear model provides one very general measure of whether treatment effects are large or small (i.e., whether they account for a lot of the variance in the dependent variable or only a little). As noted above, the value of *PV* is closely linked to *F*.

There are a number of specific statistics that are used in estimating *PV*, notably eta squared and R^2, which are typically encountered in the contexts of the analysis of variance and multiple regression, respectively. We prefer to use the more general term *PV*, because describes a general index of the effects of treatments or interventions, and is not limited to any specific statistic or statistical approach. As we will show later, estimates of *PV* are extremely useful in structuring statistical power analyses for virtually any of the specific applications of the general linear model.

Boxed section

Understanding Linear Models

Linear models combine simplicity, elegance and robustness, but for people who are the least bit math-phobic, they can seem intimidating. Consider the model illustrated in Equation 1 (presented earlier in this chapter). The mathematical form of this model is:

$y_{ijk} = a_i + b_j + ab_{ij} + e_{ijk}$

This model is easier to understand if we re-write it as a story or an explanation. This equation says:

In order to understand why some children perform better than others in reading (y_{ijk}), it is important to consider three things. First, the type of instruction (a_i) matters. Second, it makes a difference whether or not the child gets audio-visual aids (b_j). Third, these two variables might interact; the effects of methods of instruction might be different for children who receive audio-visual aids than for those who do not (ab_{ij}). Finally, there are all sorts of other variables that might be important but the effects of these variables were not measured and cannot be estimated in this experiment. The combined effects of all of the other things that affect performance introduce some error (e_{ijk}) into the explanation of why some children end up performing better than others.

Linear models are easiest to understand if you think of them as an answer to the question "Why do the scores people receive on the dependent variable (Y) vary?". All linear models answer that question by identifying some systematic sources of variance (here $a_i + b_j + ab_{ij}$); whatever cannot be explained in terms of these systematic sources is explained in terms of error (e_{ijk}).

Boxed section

The F Distribution and Power

If you take the ratio of any two independent estimates of the variance in a population (e.g., s^2_1 and s^2_2), this ratio is distributed as *F*, where:

$$F = s^2_1 / s^2_2 \qquad [2]$$

The most familiar example of the F statistic is in the analysis of variance, where s^2_1 is a measure of variability between groups and s^2_2 is a measure of variability within groups. The distribution of this *F* statistic depends of the degrees of freedom for the numerator (s^2_1) and the denominator (s^2_2); these degrees of freedom are a function of the number of pieces of information used to create s^2_1 (e.g., the number of groups) and s^2_2 (e.g., the number of subjects within each group). The *F* ratio can be used to test a wide range of statistical hypotheses (e.g., testing for the equality of means or variances). In the general linear model, the *F* statistic is used to test the null hypothesis (e.g., that means are the same across treatments), by comparing a measure of the variability in scores due to the treatments to a measure of the variability in scores you might expect as a result of sampling error. In its most general form, the *F* test in general linear models is:

$$F = \text{variability due to treatments/variability due to error} \qquad [3]$$

The distribution of the statistic *F* is complex, and it depends in part on the degrees of freedom of the hypothesis or effect being tested (df_{hyp}) and the degrees of freedom for the estimate of error used in the test (df_{err}). If treatments have no real effect, the expected value of *F* is very close to 1.0 [$E(F) = df_{err} /(df_{err} - 2)$]. That is, if the traditional null hypothesis is true the *F* ratios will usually be approximately 1.0. Sometimes, the *F* ratio might be a bit larger than 1.0 and sometimes it might be a bit smaller; depending on the degrees of freedom (df_{hyp} and df_{err}), the *F* values that are expected if the null hypothesis is true might cluster closely around 1.00, or they might vary considerably. The *F* tables shown in most statistics textbooks provide a sense of how much *F* values might vary as a function of sampling error, given various combinations of df_{hyp} and df_{err}.

The *F* and chi-squared distributions are closely related. The ratio of two chi-squared variables, each divided by its degrees of freedom, is distributed as *F*, and both distributions are special cases of a more general form (the gamma distribution). Similarly, the *t* distribution is closely related to the *F* distribution; when the means of two groups are compared, the value if *t* is identical to the square root of *F*.

<u>The *F* statistic and effect size measures.</u> The method of power analysis that this book describes is built around the *F* distribution. There are three reasons for using *F* as the centerpiece of our methods. First, as we noted earlier, most widely used statistics can be thought of as specialized applications of the general linear model, and therefore can be easily expressed in terms of *F* statistics. This means that you do not need different power tables for *F, t, r,* etc. Second, as we describe below, the well-known noncentral *F* distribution provides a simple method for evaluating power for a much wider range of hypothesis than the traditional nil hypothesis. Thus, if you want to test the hypotheses that differences between means are large enough to be worth paying attention to (rather than the hypothesis that they are exactly zero), the noncentral *F* makes this easy. Finally, it is easy to translate *F* statistics into effect size measures, in particular, into the percentage of variance in your dependent variable that is explained by the variable or variables that represent your independent variables. This effect size measure is easily applicable to any analysis involving correlation, regression, *t*-tests, analyses of variance and covariance, chi squared statistics, and many more. The anility to express so many methods of analysis in terms of the *F* statistic and to express the size of the effect any of these analyses involves is so useful and so fundamental to understanding how power analysis works and how it can be creatively applied that our first appendix (which is the one we think you will consult most often) is devoted to laying out the simple equations that allow one to translate statistical tests involving correlation,

regression, *t*-tests, analyses of covariance, chi square statistics and the like into their equivalent *F* values *and* to determine the proportion of the variance in your dependent variable that is explained in any of these analyses. Appendix A provides equations for expressing a wide range of statistics in terms of *F* and for evaluating the proportion of variance that is explained in the dependent variable.

The noncentral *F*. Most familiar statistical tests are based on the central *F* distribution - i.e., the distribution of *F* statistics you expected when the traditional nil hypothesis is true. However, interventions or treatments normally have at least some effect, and the distribution of *F* values that you actually would expect in any particular study is likely to take the form of a *singly noncentral F distribution*. The power of a statistical test is defined by the proportion of that singly noncentral *F* distribution that exceeds the critical value used to define "statistical significance".

The shape and range of values in this noncentral *F* distribution is a function of three parameters: (1) df_{hyp} (2) df_{err} and (3) the "noncentrality" parameter (λ). One way to think of the noncentrality parameter is that it is a function of just how wrong the traditional null hypothesis is. That is, the larger the difference between treatments in the populations, the larger the value of the noncentrality parameter. When $\lambda = 0$, the traditional null hypothesis is true and the noncentral *F* is identical to the central *F* that is tabled in most statistics texts.

The exact value of the noncentrality parameter is a function of both ES and the sensitivity of the statistical test (which is largely a function of the number of observations, *N*). For example, in a study where *n* subjects are randomly assigned to each of four treatment conditions, $\lambda = [n\Sigma(\mu_j - \mu)^2]/ \sigma^2_e$, where μ_j and μ represent the population mean in treatment group j and the population mean over all four treatments, and σ^2_e represents the variance in scores due to sampling error. Horton (1978) noted that in many applications of the General Linear Model:

$$\lambda_{est} = SS_{hyp} / MS_{err} \qquad [4]$$

where λ_{est} represents an estimate of the noncentrality parameter, SS_{hyp} represents the sum of squares for the effect of interest, and MS_{err} represents the mean square error term used to test hypotheses about that effect. Using *PV* to designate the proportion of the total variance in the dependent variable explained by treatments (which means that 1- *PV* refers to the proportion not explained), the noncentrality parameter can be estimated using the following equation:

$$\lambda_{est} = df_{err}\ [PV/(1- PV)] \qquad [5]$$

Equations 4 and 5 provide a practical method for estimating the value of the noncentrality parameter in a many of the most common applications of the general linear model. [1] A more general form of Equation 5 that is useful form complex analyses of variance is:

$$\lambda_{est} = [\underline{N} – p] * [PV/(1- PV)] \qquad [6]$$

Where:

N = number of subjects

p = number of terms in the linear model

The noncentrality parameter reflects the positive shift of the F distribution as the size of the effect in the population increases (Horton, 1978). For example, if N subjects are randomly assigned to one of k treatments, the mean of the noncentral F distribution is approximately $[((N-k)(k + \lambda – 1))/((k–1)\ (N–3))]$ as compared to an approximate mean of 1.0 for the central F distribution. More concretely, assume that 100 subjects are assigned to one of four treatments. If the null hypothesis is true, the expected value of F is approximately 1.0, and the value of F needed to reject the null hypothesis (α = .05) is 2.70. However, if the effect of treatments is in fact large (e.g., PV = .25), you should expect to find F values substantially larger than 1.0 most of the time; given this effect size, the mean of the noncentral F distribution is approximately 11.3, and more than 99% of the values in this distribution exceed 2.70. In other words, the power of this F test, given the large effect size of PV =.25, is greater than .99. If the population effect is this large, a statistical test comparing k groups and using a sample of N = 100 is virtually certain to correctly reject the null hypothesis.

The larger the effect, the larger the noncentrality parameter, and the larger the expected value of *F*. The larger the *F*, the more likely it is that H_0 will be rejected. Therefore, all other things being equal, the more noncentrality (i.e., the larger the effect or the larger the *N*) the higher the power.

Using the Noncentral F Distribution to Assess Power

Chapter 1 laid out the three steps in conducting a statistical power analysis, determining the critical value for significance, estimating ES, and estimating the proportion of test statistics likely to exceed critical value. Applying these three steps here, it follows that power analysis involves:

1. Deciding the value of F that is needed to reject H_0. As we will see later in this chapter, this depends in part on the specific hypothesis being tested

2. Estimating the ES and the degree of noncentrality. Estimates of *PV* can be used to estimate the noncentrality parameter of the *F* distribution

3. Estimating the proportion of the noncentral F that lies above the critical F from step 1

In the chapters that follow, we present a simple method of conducting power analyses that is based on the noncentral *F* distribution. This method can be used with a wide variety of statistics, and can be used for testing both nil hypotheses and null hypotheses that are based on specifications of negligible vs. important effects. Methods of approximating the singly noncentral *F* distribution are presented in subsequent chapters, and methods of estimating more complex noncentral *F* distributions are included in the software that is distributed with this book. Appendix B presents a table of *F* values obtained by estimating the noncentral *F* distribution over a range of df_{hyp}, df_{err}, and effect size values. Appendix C presents a parallel table that presents the percentage of variance explained (*PV*) that corresponds with each of the noncentral *F* values shown in Appendix B.

The table presented in Appendices B and C save you the difficulty of estimating noncentral *F* values, and more important, of directly computing power estimates for each statistical test. This table can be used to test both traditional and minimum-effect null hypotheses, and to estimate the statistical power of tests of both types of hypotheses.

Approximations. The methods described in this book are simple and general, but are not always precise to the last decimal place. There are many statistical procedures that fall under the umbrella of the "general linear model", and some specific applications of this model may present unique complications, or distributional complexities. However, the methods described in this book provide acceptably accurate approximations for the entire range of statistical tests covered under this general model.

Approximations are particularly appropriate for statistical power analysis because virtually all applications of this technique are themselves approximations. That is, because the exact value of the population effect size is rarely known for sure, power analyses usually depend on approximations and conventions rather than working from precise knowledge of critical parameters. Thus, in power analysis good approximations are usually quite acceptable.

For example, power analysis might be used to guide the selection of sample sizes or significance criteria in a study. Power analysis typically functions as a decision aid rather than as a precise forecasting technique, and it is rare that different decisions will be reached when exact vs. approximate power values are known. That is, users of power analysis are likely to reach the same decisions if they know that power is approximately .80 in a particular study as they would reach if they knew that the power was precisely .815.

Statistical power analysis is an area where precision is not of sufficient value to justify the use of cumbersome methods in pursuit of the last decimal place. It is possible to use the general method presented in this book to closely approximate the more specific findings that are obtained when power analyses are tailored to specific analytic techniques (see Cohen, 1988, for discussions of power analysis for each of several types of statistical tests). Our approach allows researchers to estimate statistical power for statistical tests in the general linear model, by translating specific test statistics or effect size measures into their equivalent *F* values.

Translating Common Statistics and ES Measures into F

The model developed here is expressed in terms of the *F* statistic, which is commonly reported in studies that employ analysis of variance or multiple regression. However, many studies report results in terms of statistics other than the *F* value. It is useful to have at hand formulas for translating common statistics and effect size. Appendix A presents these formulas

Suppose a study compared the effectiveness of two smoking cessation programs, using a sample of 122 adults, who were randomly assigned to treatments. The researchers used an independent-groups *t*-test to compare scores in these two treatments, and reported *t* value of 2.48. This *t* value would be equivalent to an *F* value of 6.15, with 1 and 120 degrees of freedom. The tabled value for *F* with 1 and 120 degrees of freedom (α = .05) is 3.91, and the researchers would be justified in rejecting the null hypothesis.

Suppose another study (*N* = 103) reported a squared multiple correlation of R^2 = .25 between a set of four vocational interest tests and an occupational choice measure. Applying the formula shown in Table 2-1, this R^2 value would yield an *F* value of 8.33, with 4 and 100 degrees of freedom. The critical value of *F* needed to reject the traditional null hypothesis (α = .05) is 2.46; the reported R^2 is significantly different from zero.

Suppose that in another study, hierarchical regression is used to determine the incremental contribution of several new predictor variables over and above the set of predictor variables already in an equation. For example, in a study with *N* = 250, two spatial ability tests were used to predict performance as an aircraft pilot; scores on these tests explained 14% of the variability in pilots' performance (i.e., R^2 = .14). Four tests measuring other cognitive abilities were added to the predictor battery, and this set of six tests explained an additional 15% of the variance in performance (i.e., when all six tests are used to predict performance, R^2=.29). The *F* statistic that corresponds to this increase in R^2 is *F*(4, 243) = 12.83.

Beginning of Boxed section

Worked Example – Hierarchical Regression

In Appendix A, we presented a formula that can be used to calculate the *F*-equivalent of the increase in R^2 reported in a study that used hierarchical regression. The formula is:

$$\underline{F}(\mathrm{df}_{\mathrm{hyp}}, \mathrm{df}_{\mathrm{err}}) = \frac{(\mathrm{R}^2{}_{\mathrm{F}} - \mathrm{R}^2{}_{\mathrm{R}}) / \mathrm{df}_{\mathrm{hyp}}}{(1 - \mathrm{R}^2{}_{\mathrm{F}}) / \mathrm{df}_{\mathrm{err}}}$$

In this study (N = 250), the researchers started with two predictors and reported R^2=.15, then added four more predictors and reported R^2=.29. It follows that:

$\mathrm{df}_{\mathrm{hyp}}$ = 4 this is the number of predictors that is added to the original set of two predictors. The null hypothesis is that adding these four predictors leads to no real increase in R^2

$\mathrm{df}_{\mathrm{err}}$ = 250 – 6 – 1 = 243

R^2_{F}=.29 the full model containing all six tests explains 29% of the variance in performance

R^2_{R}=.14 the restricted model that contains only the first two tests explains 14% of the variance

$F(4,243)$ = [(.29 - .14)/4] / [(1-.29)/243]

$F(4,243)$ = .037/.000292 = 12.83

Most F tables will not report critical values for 4 and 243 degrees of freedom, but if you interpolate between the critical values of F (α=.05) for 4 and 200 degrees of freedom and for 4 and 300 degrees of freedom, you will find that F values of 2.41 or larger will allow you to reject the null hypothesis. With $F(4, 243) = 12.83$, you can easily reject the null hypothesis and conclude that adding these four predictors does lead to a change in R^2.

End of Boxed section

As Appendix A shows, χ^2 values can be translated in F -equivalents. For example, if a researchers found a χ^2 value of 24.56 with 6 degrees of freedom, the equivalent F value would be 4.09 (i.e., 24.56/6), with $\mathrm{df}_{\mathrm{hyp}}$ =6 and $\mathrm{df}_{\mathrm{err}}$ being infinite. Because the F table asymptotes as $\mathrm{df}_{\mathrm{err}}$ grows larger, $\mathrm{df}_{\mathrm{err}}$ =10,000 (which is included in the F Table listed in C) represents an excellent approximation to infinite degrees of freedom for the error term. The critical value of F (α=.05) needed to reject the null hypothesis in this study is 2.10, so once again, the null hypothesis will be rejected.

Appendix A also includes the ES measure *d*. This statistic is not commonly used in hypothesis testing per se, but it is widely used in describing the strength of effects, particularly when the scores of those receiving a treatment are compared to scores in a control group. This statistic can also be easily transformed into its *F* -equivalent using the formulas shown in Appendix A.

Finally, a note concerning terminology. In the section above, and in several sections that follow, we use the term "*F*-equivalent". We this term to be explicit in recognizing that even when the results of a statistical test in the general linear model are reported in terms of some statistic other than *F*, (e.g., *r, t, d*), it is nevertheless usually possible to transform these statistics into the *F* value that is equivalent in meaning.

Beginning of Boxed Section

<u>Worked Examples Using the *d* Statistic</u>

The *d* statistic can be used to describe the strength of an effect in a single study. For example, if a researcher was comparing two treatments and reported *d* = .50, this would indicate that the difference between treatment means was one half as large as the standard deviation within each group. An even more common use of *d* is in meta-analysis, in which the results of several studies are summarized to estimate how large an effect is in the population.

Suppose, for example, that previous research suggests that the effect size *d* should be about .25. This ES measure would be extremely useful in conducting power analyses. Given this value of *d*, a study in which 102 subjects were randomly assigned to one of these two treatments would be expected to yield:

$$F(1, \mathrm{df}_{\mathrm{err}}) = \frac{d^2 \mathrm{df}_{\mathrm{err}}}{4}$$

Where:

d = .25

$\mathrm{df}_{\mathrm{err}}$ = N – 2 = 102 – 2 = 100

and

F(1,100) = [$.25^2$ * 100]/4

F(1,100) = [.0625 * 100] / 4 = 1.56

If you check an *F* table, you will find that this *F* is not close to the value needed to reject a null hypothesis (for df=1,100, you need an F value of 3.93 to reject the null hypothesis). That is given the effect size expected here, a sample of *N* = 102 will not provide very much statistical power.

The calculations above are based on the independent-groups *t*, in which participants are randomly assigned to one of two treatments. Suppose you used a repeated measures design (e.g., one in which scores of 102 subjects on a pretest and a posttest were compared, with a correlation of .60 between these scores). If you expect that *d* =.25, the formula for transforming repeated-measures t yields:

$$F(1, \mathrm{df}_{\mathrm{err}}) = \frac{d^2 \mathrm{df}_{\mathrm{err}}}{4\sqrt{1 - r_{ab}}}$$

Where:

d = .25

$\mathrm{df}_{\mathrm{err}}$ = N – 2 = 102 – 2 = 100

r_{ab} = .60

and:

F(1,100) = [$.25^2$ * 100]/[4 * .632]   [the square root of (1-.60) is .632]

F(1,100) = [.0625 * 100] / 2.529 = 2.47

The critical value for *F* (α = .05) when the degrees of freedom are 1 and 100 is 3.93, suggesting that the studies described above would not have sufficient power to allow you to reject the traditional null hypothesis.

As you may have noticed, in the examples above we included the sample size. The reason for this is that the value and the interpretation of the *F* statistic depends in part on the size of the sample (in particular, on the degrees of freedom for the error term, or $\mathrm{df}_{\mathrm{err}}$). In the preceding paragraph, a *d* value of .25 in a sample of 102 would yield *F* (1,100) = 1.56. In a study that randomly assigned 375 participants to treatments, the same *d* would translate into *F* (1,373) = 5.82, which would be statistically significant. This reflects the fact that the same difference between means is easier to statistically detect when the sample (and therefore $\mathrm{df}_{\mathrm{err}}$) is large than when the sample is small. Small samples produce unstable and unreliable results, and in a small sample it can be hard to distinguish between true treatment effects and simple sampling error.

End of Boxed Section

Transforming from F to PV . Appendix A shows how to transform commonly used statistics and effect size estimates into their equivalent *F* values. It can also often be used transform from *F* values into ES measures. For example, suppose you randomly assigned participants into four different treatments and used analysis of variance to analyze the data. You reported a significant *F* value [F(3, 60) = 2.80], but this does not provide information about the strength of the effect. Formula 6 allows you to obtain an estimate of the proportion of variance in the dependent variable explained by the linear model (i.e., *PV*), given the value of *F*:

$$PV = (\mathrm{df}_{\mathrm{hyp}} * F) / [(\mathrm{df}_{\mathrm{hyp}} * F) + \mathrm{df}_{\mathrm{err}}] \qquad [7]$$

While yields:

$$PV = (3 * 2.80) / [(3 * 2.80) + 60] \qquad [8]$$

$$PV = 8.4 / 68.4 = .122$$

In other words, the *F* value reported in this study allows you to determine that treatments accounted for 12% of the variance in outcomes.

Formula 6 cannot be used in complex, multifactor analyses of variance, because the *F* statistic for any particular effect in a complex ANOVA model and its degrees of freedom do not contain all of the information needed to estimate *PV*. In chapters 7 and 8, we will discuss the application of power analyses to these more complex designs, and will show how information presented in significance tests can be used to estimate ES.

Defining Large, Medium and Small Effects

Cohen's books and papers on statistical power analyses (e.g., Cohen, 1988) have suggested a number of conventions for describing treatment effects as "small", "medium", or "large". These conventions are based on surveys of the literature, and seem to be widely accepted, at least as approximations. Table 2-1 presents conventional values for describing large, medium, and small effects, expressing these effects in terms of a number of widely-used statistics.

Insert Table 2-1 about here

For example, a small effect might be described as one that accounts for about 1% of the variance in outcomes, or one where the treatment mean is about one fifth of a standard deviation higher in the treatment group than in the control group, or as one where the probability that a randomly selected member of the treatment group will have a higher score than a randomly selected member of the control group is about .56.

These are, of course, not the only suggestions for describing effects as small, medium or large. For example, Ferguson (2009) describes a correlation of .20 or a

Table 2-1

Some Conventions for Defining Effect Sizes

	PV	r	d	f^2	probability of a higher score in treatment group
Small effects	.01	.10	.20	.02	.56
Medium effects	.10	.30	.50	.15	.64
large effects	.25	.50	.80	.35	.71

From: Cohen (1988), Grissom (1994), Ferguson (2009)

Note: Cohen's $f^2 = R^2/(1- R^2) = \eta^2/(1- \eta^2) = PV/(1-PV)$, where

$\eta^2 = SS_{treatments}/SS_{total}$

squared correlation (PV) of .04 as the smallest value that is likely to be practically significant. Our preference is to use the more conservative threshold suggested by Cohen (i.e., PV = .01 or lower) to describe effects that are likely to be too small to matter in most real-world settings.

The values in Table 2-2 are approximations and nothing more. In fact, a few minutes with a calculator shows that they are not all exactly equivalent (e.g., if you square an *r* value of .30, you get an estimate of *PV* =.09, not *PV* = .10). Although they are not exact or completely consistent, the values in Table 2-2 are nevertheless very useful. These conventions provide a starting point for statistical power analysis, and they provide a sensible basis for comparing the results in any one study with a more general set of conventions.

Nonparametric and Robust Statistics

The decision to anchor our model for statistical power analysis to the *F* distribution is driven primarily by the widespread use of statistical tests based on that distribution. The *F* statistic can be used to test virtually any hypothesis that falls under the broad umbrella of the "general linear model". There are, however, some important statistics that do not fall under this umbrella, and these are not easily transformed into a form that is compatible with the *F* distribution.

For example, a number of robust or "trimmed" statistics have been developed in which outliers are removed from observed distributions prior to estimating standard errors and test statistics (Wilcox, 1992; Yuen, 1974). Trimming outliers from data can sometimes substantially reduce the effects of sampling error, and trimmed statistics can have more power than their normal-theory equivalents (Wilcox, 1992). The power tables developed in this book are not fully appropriate for trimmed statistics, and can substantially underestimate the power of these statistics when applied in small samples.

A second family of statistics that are not easily accommodated using the model developed here are those statistics referred to as "nonparametric" or distribution-free statistics. Nonparametric statistics do not require a priori assumptions about distributional forms, and tend to use little information about the observed distribution of data in constructing statistical tests. The conventional wisdom has long been that nonparametric tests have less power than their parametric equivalents (Siegel, 1956), but this is not always the case. Nonparametric tests can have more power than their parametric equivalents under a variety of circumstances, especially when conducting tests using distributions with heavy tails (i.e., more extreme scores than would be expected in a normal distribution; Zimmerman & Zumbo, 1993). The methods developed here do not provide accurate estimates of the power of robust or nonparametric statistics.

From F to Power Analysis

Earlier in this chapter, we noted that statistical power analysis involves a three-step process. First, you must determine the value of *F* that is needed to reject the null hypothesis – i.e., the critical value of *F*. Virtually any statistics text is likely to include a table of critical *F* values that can be used to test the traditional null hypothesis. The great advantage of framing statistical power analysis in terms of the noncentral *F* distribution is that this approach makes it easy to test a number of alternatives to the nil hypothesis. As we will show in chapter 3 this approach to statistical power analysis makes it easy to evaluate the power of tests of the hypothesis that treatments have effects that are not only greater than zero, but that are also sufficiently large that they are substantively meaningful. Like tests of the traditional null hypothesis, these minimum-effect tests start with the identification of critical values for the *F* statistic.

Once the critical value of *F* is established for any particular test, the assessment of statistical power is relatively easy. As we noted in Chapter 1, *the power of a statistical test is the proportion of the distribution of test statistics expected for a particular study that is above the critical value used to establish statistical significance.* If you determine that the critical value for the *F* statistic that will be used to test hypotheses in your study is equal to 6.50, all you need to do to conduct a power analysis is to determine the noncentral *F* distribution that corresponds with the design of your study (the design of your study determines df_{hyp} and df_{err}) and the ES you expect in that study, which determines the degree of noncentrality of the *F* distribution. If the critical value of *F* is equal to 6.50, power will be defined as the proportion of this noncentral *F* distribution that is equal to or greater than 6.50.

A number of methods can be used to carry out statistical power analyses. For the mathematically inclined, it is always possible to analytically estimate the noncentral *F* distribution; we will discuss analytic methods below. However, most users of power analysis are likely to want a simpler set of methods. We will discuss the use of power tables and power calculators in the section below.

Analytic and Tabular Methods of Power Analysis

Analytic methods of power analysis are the most flexible and the most exact, but also the most difficult to implement. Power tables or graphs provide good approximations of the statistical power of studies under a wide range of conditions. Our preference is to work with tables, in part the type of graphs needed to plot a variable (i.e., power) as a function of three other variables (i.e., *N*, α, and ES) strike us as complicated and difficult to use. Software and calculators that can be used to estimate statistical power combine the flexibility and precision of analytic methods with the ease of use of power tables.

Analytic methods. The most general method for evaluating statistical power involves estimating the noncentral *F* distribution that corresponds with the design of your study and the ES you expect. While this analytic method is both precise and flexible, it is also relatively cumbersome and time-consuming. That is, the direct computation of statistical power involves: (1) determining some standard for statistical or practical significance (2) estimating the noncentral *F* distribution that corresponds to the statistic and study being analyzed, and (3) determining the proportion of that noncentral *F* distribution that lies above the standard. Even with a relatively powerful computer, the processes can time-consuming, and may be daunting to many consumers of power analysis; for readers interested in analytic approaches, Appendix D presents several methods for estimating the necessary distributions. A more user-friendly approach is to develop tables or calculators that contain the essential information needed to estimate statistical power.

Power tables. A number of excellent books present extensive tables describing the statistical power of numerous tests; Cohen (1988) is the most complete source currently available. The approach presented here is simpler (although it provides a bit less information), but considerably more compact. Unlike Cohen (1988), who developed tables for many different statistics, our approach involves translating different statistics into their *F* -equivalents. This allows us to present virtually all of the information needed to do significance tests and power analyses for statistical tests in the general linear model in a single table.

Appendix B contains a table we call the "One Stop *F* Table". This is called a "one stop" table because each cell contains the information you need for: (1) conducting traditional significance tests, (2) conducting power analyses at various key levels of power, (3) testing the hypothesis that the effect in a study exceeds various criteria used to define negligibly small or small to moderate effects, and (4) estimating power for these "minimum-effect" tests. Minimum-effects tests are explained in detail in Chapter 3; at this point we will focus on the use of the One Stop *F* Table for testing the traditional null hypothesis and for evaluating the power of these tests.

Using the One-Stop F Table

Each cell in the One-Stop *F* Table contains twelve pieces of information. The first four values in each cell are used for testing significance and estimating power for traditional null hypothesis tests. The next eight values in each cell are used for testing significance and estimating power when testing the hypothesis that treatment effects are negligible, using two different operational definitions of a "negligible" effect (i.e., treatments account for 1% or less of the variance, or that they account for 5% or less of the variance). In this chapter, we will focus on the first four pieces of information presented for each combination of df_{hyp} and df_{err}.

This table presents the critical value of *F* for testing the traditional null hypothesis, using α values of .05 and .01, respectively. We label these “nil .05” and “nil .01”. For example, consider a study in which fifty-four subjects are randomly assigned to one of four treatments, and the analysis of variance is used to analyze the data. This study will have df_{hyp} and df_{err} values of 3 and 50, respectively. The critical values of *F* for testing the nil hypothesis will be 2.79 when α equals .05; the critical value of *F* for testing the nil hypothesis will 4.20 when α equals .01. In other words, in order to reject the nil hypothesis ($\alpha = .05$), the value of the mean square for treatments will have to be at least 2.79 times as large as the value of the mean square for error (i.e., $F = MS_{treatments} / MS_{error}$).

The next two values in each cell are *F* - equivalents of the effect size values needed to obtain power of .50 and power of .80 (we label these “pow .50” and “pow .80”), given an α level of .05 and the specified df_{hyp} and df_{err}. The values of “pow .50” and “pow .80” for 3 and 50 degrees of freedom are in the table are 1.99 and 3.88, respectively. That is, a study designed with 3 and 50 degrees of freedom will have power of .50 for detecting an effect that is equivalent to $F = 1.99$. It will have power of .80 for detecting an effect that is equivalent to $F = 3.88$. *F* -equivalents are handy for creating tables, but in order to interpret these values, it is necessary to translate them into ES measures.

Because subjects are randomly assigned to treatments and the simple analysis of variance is used to analyze the data, we can use Formula 6, presented earlier in this chapter, to transform these *F* values into equivalents *F* values (i.e., $PV = (df_{hyp} * F) /$

[(df_{hyp} * F) + df_{err}]). Applying this formula, you will find that in a study with df_{hyp} and df_{err} values of 3 and 50, you will need a moderately large ES to achieve power of .50. Translating the F value of 2.16 into its equivalent *PV*, you will find:

PV = (3 * 1.99) / [(3 * 1.99) + 50] = 5.97/55.97 = .106

That is, in order to achieve a power of .50 with df_{hyp} and df_{err} values of 3 and 50, you will need to be studying treatments that account for at least 10% of the variance in outcomes. Applying the same formula to the F needed to achieve power of .80 (i.e., F = 3.88), you will find that you need an effect size of PV = .188. In other words, in order to achieve power of .80 with this study, you will need to be studying a truly large effect, on in which treatments account for about 19% or more of the variance in outcomes.

Suppose that on the basis of your knowledge of the scientific literature, you expect the treatments being studied to account for about 15% of the variance in outcomes. This ES falls between PV = .10 and PV = .19, which implies that power of this study will fall somewhere between .50 and .80. As we show in the section that follows, it is easy to estimate where in this range the power of this study actually falls (in this example, power is approximately .65).

Interpolating between tabled values. Like all *F* tables, our "One-Stop *F* Table" is incomplete, in that it does not table all possible values of df_{hyp} and df_{err}. Fortunately, relatively good approximations to all of the values in this table can be obtained by linear interpolation. For example, our table includes df_{err} values of 50 and 60. If you wanted to find appropriate F values for df_{err} =55, these would lie about halfway between the values for df_{err} =50 and df_{err} =60. Thus, the approximate F needed to reject the traditional null hypothesis (α = .05) with df_{hyp} =2, df_{err} =55 would be 3.165 (i.e., halfway between 3.15 and 3.18. Similarly, if df_{err} =48, you could estimate the appropriate *F* values by computing the value that *F* that was 80% of the distance between the tabled *F* for df_{err} =40 and the tabled *F* for df_{err} =50. In general, the value of the interpolated *F* can be obtained using the formula below:

$$F_{\text{interpolated}} = F_{\text{below}} + \frac{(df_{\text{int}} - df_{\text{below}})}{(df_{\text{above}} - df_{\text{below}})}(F_{\text{above}} - F_{\text{below}}) \qquad [9]$$

Where:

F_{below} = Tabled *F* below the value to be calculated

F_{above} = Tabled *F* above the value to be calculated

df_{below} = df_{err} for tabled *F* below the value to be calculated

df_{above} = df_{err} for tabled *F* above the value to be calculated

df_{int} = df_{err} for *F* value to be calculated

It is important to keep in mind that linear interpolation will yield approximate values only. For the purposes of statistical power analyses, these interpolations will virtually always be sufficiently accurate to help you to make sensible decisions about the design of studies, the choice of criteria for defining "statistical significance", etc.

Beginning of Boxed section

Worked Example – Interpolating Between Values for Power of .50 and .80

A second application of linear interpolation is likely to be even more useful. Our table includes *F* - equivalents for the effect size values needed to obtain power levels of .50 and .80, respectively. In the example described earlier, where df_{hyp} =3 and df_{err} =50, F values of 1.99 and 3.88 are equivalent to the population *PV* values that would be needed to obtain power levels of .50 or .80. These *F* values translate in *PV* values of in a study like this .10 and .19, respectively. Here, you expected treatments to account for 15% of the variance. If you translate this figure into its *F* -equivalent (using formulas presented in Table 2-1), we will obtain *F* of 2.94. You can use linear interpolation to estimate the power of this study, using a formula that closely parallels the formula used to interpolate *F* values:

$$\text{Power}_{\text{interpolated}} = .50 + \left[\left(\frac{F_{\text{hypothesized}} - F_{.50}}{F_{.80} - F_{.50}}\right) * .30\right] \qquad [10]$$

Where:

$F_{\text{hypothesized}}$ = *F* - equivalent for hypothesized size of the effect

$F_{.50}$ = *F* -equivalent of the *PV* needed to obtain power of .50 (α = .05)

$F_{.80}$ = *F* -equivalent of the *PV* needed to obtain power of .80 (α = .05)

In the example discussed above df_{hyp} =3 and df_{err} =50. If you expect treatments to account for 15% of the variance, the equivalent *F* value is 2.94. In terms of equation 10, $F_{\text{hypothesized}}$ = 2.94, $F_{.50}$ = 1.99, and $F_{.80}$ = 3.88, which means that the power of this study for rejecting the null hypothesis (α = .05) is .64.

End of Boxed section

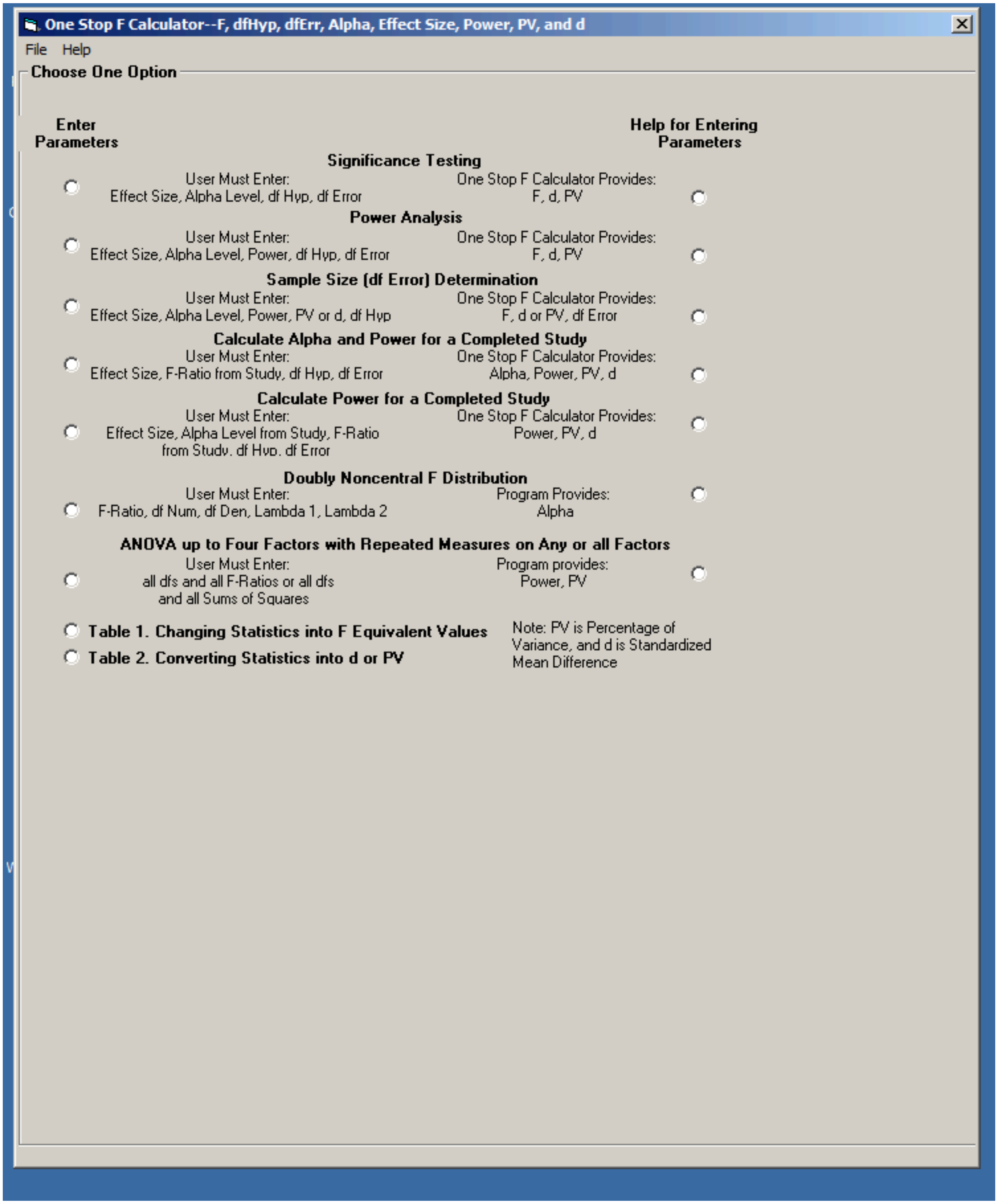
The One Stop F Calculator

It is often easier and more convenient to use a computer program rather than a set of tables to estimate statistical *power*. The One Stop *F* Calculator program distributed with this book is the computer program analog to all tables at the back of this book, including the One Stop *F* Table and the One Stop *PV* Table (Appendices B and C). The One Stop *F* Calculator program is written in Visual Basic and uses subroutines based on programs developed by Reeve (1986a, 1986b, 1986c).

Figure 2.1 illustrates the screen that opens when the One Stop *F* Calculator is installed on a computer. This calculator contains five general options for power

Figure 2-1

One-Stop F Calculator



One Stop F Calculator--F, dfHyp, dfErr, Alpha, Effect Size, Power, PV, and d
File
Help

Choose One Option

Enter Parameters

Help for Entering Parameters

Significance Testing

○ User Must Enter: Effect Size, Alpha Level, df Hyp, df Error
One Stop F Calculator Provides: F, d, PV ○

Power Analysis

○ User Must Enter: Effect Size, Alpha Level, Power, df Hyp, df Error
One Stop F Calculator Provides: F, d, PV ○

Sample Size (df Error) Determination

○ User Must Enter: Effect Size, Alpha Level, Power, PV or d, df Hyp
One Stop F Calculator Provides: F, d or PV, df Error ○

Calculate Alpha and Power for a Completed Study

○ User Must Enter: Effect Size, F-Ratio from Study, df Hyp, df Error
One Stop F Calculator Provides: Alpha, Power, PV, d ○

Calculate Power for a Completed Study

○ User Must Enter: Effect Size, Alpha Level from Study, F-Ratio from Study, df Hyp, df Error
One Stop F Calculator Provides: Power, PV, d ○

Doubly Noncentral F Distribution

○ User Must Enter: F-Ratio, df Num, df Den, Lambda 1, Lambda 2
Program Provides: Alpha ○

ANOVA up to Four Factors with Repeated Measures on Any or all Factors

○ User Must Enter: all dfs and all F-Ratios or all dfs and all Sums of Squares
Program provides: Power, PV ○

○ **Table 1. Changing Statistics into F Equivalent Values**

○ **Table 2. Converting Statistics into d or PV**

Note: PV is Percentage of Variance, and d is Standardized Mean Difference analysis: Significance Testing, Power Analysis, Sample Size (df Error) Determination, Calculate Alpha and Power for a Completed Study, and Calculate Power for a Completed Study. It also includes an option for computing power in multifactor experiments with up to four factors, and with repeated measures on any or all of those factors. It also includes an option for calculating the *p* value in a noncentral *F* distribution for a wide range of values of *F*, df and noncentrality parameters.

The calculator allows you to directly enter the information needed to complete each type of power analysis, or to go to a help screen that explains what information is needed and why. The calculator also gives you a quick way of looking up the various methods used to translate test statistics into their equivalent *F* and *PV* values. To use the calculator, simply select the option you want to explore, enter the requested information in the white text boxes and hit “Calculate”.

To use this calculator, several decisions must be made. First, what hypothesis is being tested? The most common statistical test involved the nil hypothesis – i.e., the hypothesis that treatments have no effect. To test the traditional null hypothesis, enter the null hypothesis being tested in that treatments account for some negligible amount of variance (e.g., 1% of the variance or less), versus the alternate hypothesis that treatments have effects that are large enough to care about. These hypotheses can be easily tested by entering the appropriate effect size (e.g., 0 for a traditional nil hypothesis, .01 for 1% of the variance, .05 for 5%) in the Effect Size text box.

Depending on the option you choose (e.g., Significance Testing), you may need to enter the alpha level, df_{hyp}, df_{err}, *PV, d,* or *F*. For example, suppose 120 subjects are randomly assigned to one of four treatments. You are interested in significance testing, using an alpha level of .05. In the "Significance Testing" section, enter values of 0, .05, 3, and 116 in the Effect Size, Alpha, df_{hyp} and df_{err} boxes and press "Calculate". The One-Stop F calculator will show that you will need an *F* value of 2.68 to achieve statistical significance, which is equivalent to a *PV* of .064.

To determine the effect size needed to achieve power of .80 for this same study, go to the "Power Analysis" option and enter values of 0, .80, .05, 3 and 116 in the Effect Size, Power, Alpha, df_{hyp} and df_{err} boxes and and press "Calculate". The One-Stop *F* calculator will show that you will need a moderately large effect size (*PV* = .087) to achieve this level of power.

Suppose you expected a somewhat smaller ES (e.g., *PV* = .05), and you want to determine the sample size needed to achieve power of .80. Go to the "Sample Size (df_{err}) Determination" option and enter values of 0, .80, .05, 3 and .05 in the Effect Size, Power, Alpha, df_{hyp} and *PV* boxes and press "Calculate". The One-Stop *F* calculator will show that you will need a df_{err} of 203 to achieve this level of power. In the oneway analysis of variance, the total sample size is given by $N = df_{hyp} + df_{err} + 1$, which means that you will need a sample of 207 to achieve power of .80.

Suppose you calculate the value of *F* and find *F* = 2.50. The "Calculate Alpha and Power for a Completed Study" option allows you to determine the confidence with which you could reject the null hypothesis. Enter values of 0, .2.50, 3 and 116 in the Effect Size, F, , df_{hyp} and df_{err} boxes and and press "Calculate". The One-Stop *F* calculator shows you could reject the traditional null hypothesis at the .06 level with power of .638. In many cases, it is difficult to convince researchers to use alpha levels other than the traditional values of .05 or .01. If you select the "Calculate Power for a

Completed Study with a Selected Alpha Level" option and enter values of 0, .05, 2.5, 3 and 116 in, and you will find that power in this study is .600 when alpha is .05.

The "Help" menu for the One-Stop *F* calculator provides a primer on power analyses, references to relevant articles and books and a discussion of the information needed to implement each of the options, as well as a discussion of the meaning of the results provided under each of the options

Summary

The statistics that are most widely used in the social and behavioral sciences are either interpreted in terms of, or easily translated into the *F* statistic. Our model for power analysis uses the noncentral *F* distribution to estimate the power of a wide range of statistics (cf. Patnaik, 1949). The distinction between central and noncentral F distributions is one between what you hypothesize in order to create a null hypothesis test and what you really expect to find in a study. In this sense noncentral *F* represents the distribution of outcomes you expect to find in any particular study (given an effect size, df_{hyp} and df_{err}); the degree of noncentrality (λ) is a direct function of the effect size of interest. The statistical power of your study is simply the proportion of this noncentral *F* distribution that lies above whatever criterion you use to define "statistical significance". This model of power analysis is not limited to tests of the traditional null hypothesis (i.e., that treatments had no effect whatsoever), but rather can be easily generalized to tests of substantively meaningful hypotheses (e.g., that the treatment effect exceeds some specific value).

We discuss both analytic and tabular methods of statistical power analysis. In particular, we introduce the "One-Stop *F* Table", which contains all of the information needed to test the null hypothesis and estimate statistical power. We also discuss a parallel table that lists *PV* values that are equivalent to the *F* s shown in the One-Stop *F* Table. This table allows researchers to conduct analyses in terms of effect size estimates rather than in terms of their *F*-equivalents. Both tables contain the same basic information, but different users might find one form or the other more convenient to use. Finally, we discuss the “One-Stop *F* Calculator” program, which is included with this book. This program gives you a simple interface for carrying out virtually all of the analyses discussed in this book.

Footnotes

[1] Equations 4 and 5 are based on simple linear models, in which there is only one effect being tested, and the variance in scores is assumed to be due to either the effects of treatments or to error (e.g., this is the model that underlies the t-test or the one-way analysis of variance). In more complex linear models, df_{err} does not necessarily refer to the degrees of freedom associated with variability in scores of individuals who receive the same treatment (within-cell variability in the one-way ANOVA model), and a more general form of Equation 5 ($\lambda_{est} = [(N-k) * (PV/(1- PV))]$, where *N* represents the number of observations and $\underline{k}$ represents the total number of terms in the linear model) is needed. When *N* is large, Equation 5 yields very similar results to those of the more general form shown above.